REMARKS

A substitute specification is submitted herewith (in both clean and marked versions), making purely formal changes to the specification. No new matter has been added.

In view of the foregoing amendments and remarks, Applicants again respectfully request favorable reconsideration and passage to issue of the present application.

Applicants' undersigned attorney may be reached in our New York office by telephone at (212) 218-2100. All correspondence should continue to be directed to our below listed address.

Respectfully submitted,

P. Diana
Attorney for Applicants

Registration No. 29,296

FITZPATRICK, CELLA, HARPER & SCINTO
30 Rockefeller Plaza
New York, New York 10112-3801
Facsimile: (212) 218-2200

NY_MAIN 393813v2

TITLE

APPARATUS AND METHOD FOR DIVIDING DOCUMENT INCLUDING TABLE

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present invention relates to document segmentation apparatus and methods for dividing a document from content to content, and more particularly it relates to document segmentation apparatus and methods for dividing a document including a table or tables.

Related Background Art

[0002] In the past, information on [[a]] the web has been presented ~~as unit of~~ in units termed "pages", and the arrangement and dimension of the page can freely be set by the information presenter. Of course, the information presenter forms the pages on the basis of his or her information transmitting intention, but[[,]] it is not ~~necessary~~ necessarily the case that such pages meet [[a]] the requirements of a reader.

[0003] Accordingly, even when a series of topics or subjects which are judged to have close relation by the presenter are gathered in one page, the reader may not want such relation, and, if only one of plural subjects is useful, information [[of]] about the other subjects may be an obstacle when required information is retrieved. Particularly, in mobile equipment[[s]] having an information presenting space, a function for displaying only required information is important.

[0004] Thus, it is important that documents to be displayed are divided into segments based on ~~from~~ content ~~to content~~ (segmentation) in advance and that only a portion which is requested by the reader ~~can be~~ is presented. In almost all [[of]] web pages, contents are written by using Hyper Text Markup Language (HTML), which is a language ~~to compose~~ for use in composing web pages. Although ~~the~~ HTML is a language for describing the structure of the document, it is difficult to describe details of theoretical structure by using ~~the~~ HTML, and [[a]] the main role of ~~the~~ HTML is to designate [[an]] the layout in the browser.

[0005] However, it is considered that the viewpoint of the information presenter is reflected [[to]] in the layout of the page. Thus, there has been proposed a technique in which the page is divided on the basis ~~of tags~~ of HTML tags in order to generate segments which reflect the intention of the information presenter.

[0006] In such a technique, a table, in the sense of a portion between the <TABLE> tag and the </TABLE> tag, is judged as one meaningful group and is formed as one segment. However, ~~the~~ such a table frequently includes a plurality of sets of information which ~~assume~~ occupy a relatively great space.

[0007] Further, ~~the~~ such "tables" can be categorized into tables in the general meaning of that word, [[or]] and table formatting used for designating the layout of image or text. In ~~both bases;~~ the two cases, tags are used in quite different ways.

[0008] Furthermore, when the table formatting describes ~~the simple~~ an actual table, a set of data is represented in a column or in a row, or there is a column (or row) with a given item name or not; ~~namely~~ that is, the table has various styles.

SUMMARY OF THE INVENTION

[0009] An object of the present invention is to divide a table into a plurality of segments on the basis of contents thereof.

[0010] Another object of the present invention is to provide a table in a document divided into segments differing from each other in ~~from content to~~ content, by analyzing the table to be processed to judge whether the table is ~~a table describing~~ ~~a~~ an actual table ~~in general meaning~~ or [[a]] table formatting being used as a tool of layout, and by generating segments accordingly.

[0011] A further object of the present invention is to provide an actual table ~~in general meaning~~ into data segments on the basis of the style of the actual table ~~in general meaning when the table describes a table in general meaning~~.

[0012] A still further object of the present invention is to generate segments on the basis of groups of contents when [[a]] table formatting is used to obtain a desired layout of image or text.

[0013] According to one aspect, the present invention which achieves these objectives relates to a document segmentation apparatus comprising table analyzing means for generating cell position data indicating a positional relationship between cells and cell vectors representing characteristics of the cells, by analyzing a table in a document to be processed, table type judging means for judging a table type with reference to the cell position data and the cell vectors generated by the table analyzing means, first segment generating means for generating a segment from the table when the table type is a table describing a table, and second segment generating means for generating a segment from the table when the table type is a table for layout.

[0014] According to another aspect, the present invention which achieves these objectives relates to a document segmentation method comprising a table analyzing step for generating cell position data indicating a positional relationship between cells and cell vectors representing characteristics of the cells, by analyzing a table in a document to be processed, a table type judging step for judging a table type with reference to the cell position data and the cell vectors generated by the table analyzing step, a first segment generating step for generating a segment from the table when the table type is a table describing a table, and a second segment generating step for generating a segment from the table when the table type is a table for layout.

[0015] According to still another aspect, the present invention which achieves these objectives relates to a computer-readable storage medium storing a document segmentation program for controlling a computer to perform document segmentation, the program comprising codes for causing the computer to perform a table analyzing step for generating cell position data indicating a positional relationship between cells and cell vectors representing characteristics of the cells, by analyzing a table in a document to be processed, a table type judging step for judging a table type with reference to the cell position data and the cell vectors generated by the table analyzing step, a first segment generating step for generating a segment from the table when the table type is a table describing a table, and a second segment generating step for generating a segment from the table when the table type is a table for layout.

[0016] Other objectives and advantages besides those discussed above ~~shall~~ will be apparent to those skilled in the art from the description of [[a]] the preferred embodiments of the intention which follows. In the description, reference is made to accompanying drawings, which form a part thereof, and which illustrate examples of the invention. Such examples, however, are not exhaustive of the various embodiments of the inventions, and therefore reference is made to claims which follow the description for determining the scope of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] Fig. 1 is a block diagram showing a functional construction of a document segmentation apparatus according to a first embodiment of the present invention;

[0018] Fig. 2 is a block diagram showing a hardware construction of the document segmentation apparatus according to the first embodiment;

[0019] Fig. 3 is a flow chart showing a procedure of the document segmentation processing according to the first embodiment;

[0020] Fig. 4 is a view for explaining a maximum distance algorithm;

[0021] Fig. 5 is a block diagram showing a functional construction according to a second embodiment of the present invention;

[0022] Fig. 6 is a block diagram showing a functional construction according to a third embodiment of the present invention;

[0023] Fig. 7 is a block diagram showing a functional construction according to a fourth embodiment of the present invention;

[0024] Fig. 8 is a view showing an example of a table in an HTML document;

[0025] Fig. 9 is a block diagram showing a functional construction according to a fifth embodiment of the present invention;

[0026] Fig. 10 is a block diagram showing a construction of a table type ~~judgement~~ judgment part according to the fifth embodiment;

[0027] Fig. 11 is a flow chart showing a procedure of a table type ~~judgement~~ judgment processing according to the fifth embodiment;

[0028] Fig. 12 is a view showing an example of a table in an HTML document;

[0029] Fig. 13 is a block diagram showing a construction of a table type ~~judgement~~ judgment part according to a sixth embodiment of the present invention;

[0030] Fig. 14 is a flow chart showing a procedure of a table type ~~judgement~~ judgment processing according to the sixth embodiment;

[0031] Fig. 15 is a view showing an example of a table in an HTML document;

[0032] Fig. 16 is a block diagram showing a construction of a table type ~~judgement~~ judgment part according to a seventh embodiment of the present invention;

[0033] Fig. 17 is a flow chart showing a procedure of a table type ~~judgement~~ judgment processing according to the seventh embodiment;

[0034] Fig. 18 is a block diagram showing a construction of a table type ~~judgement~~ judgment part according to an eighth embodiment of the present invention;

[0035] Fig. 19 is a flow chart showing a procedure of a table type ~~judgement~~ judgment processing according to the eighth embodiment;

[0036] Fig. 20 is a block diagram showing a construction of a table type ~~judgement~~ judgment part according to a ninth embodiment of the present invention;

[0037] Fig. 21 is a flow chart showing a procedure of a table type ~~judgement~~ judgment processing according to the ninth embodiment;

[0038] Fig. 22 is a block diagram showing a construction of a table type ~~judgement~~ judgment part according to a tenth embodiment of the present invention;

[0039] Fig. 23 is a flow chart showing a procedure of a table type ~~judgement~~ judgment processing according to the tenth embodiment;

[0040] Fig. 24 is a view showing an example of a table in an HTML document;

[0041] Fig. 25 is a block diagram showing a functional construction of a document segmentation apparatus according to an eleventh embodiment of the present invention;

[0042] Fig. 26 is a flow chart showing a procedure of the document segmentation processing according to the eleventh embodiment;

[0043] Fig. 27 is a flow chart showing a procedure for HTML table reformation according to the eleventh embodiment;

[0044] Fig. 28 is a view showing an example of a table in an HTML document;

[0045] Figs. 29A, 29B, 29C, 29D and 29E are flow charts showing a procedure for HTML table reformation according to a twelfth embodiment of the present invention;

[0046] Figs. 30A, 30B, 30C, 30D, 30E and 30F are views showing example of multi-row/multi-column tables;

[0047] Figs. 31A, 31B, 31C and 31D are flow charts showing a procedure for HTML table reformation according to a thirteenth embodiment of the present invention;

[0048] Figs. 32A, 32B and 32C are views showing an example of a composite table;

[0049] Fig. 33 is a block diagram showing a construction of an HTML table reformation part according to a fourteenth embodiment of the present invention;

[0050] Fig. 34 is a flow chart showing a procedure of the HTML table reformation processing according to the fourteenth embodiment;

[0051] Fig. 35 is a block diagram showing a construction of an HTML table reformation part according to a fifteenth embodiment of the present invention;

[0052] Fig. 36 is a flow chart showing a procedure of the HTML table reformation processing according to the fifteenth embodiment;

[0053] Fig. 37 is a block diagram showing a construction of an HTML table reformation part according to a sixteenth embodiment of the present invention;

[0054] Fig. 38 is a flow chart showing a procedure of the HTML table reformation processing according to the sixteenth embodiment;

[0055] Fig. 39 is a block diagram showing a construction of an HTML table reformation part according to a seventeenth embodiment of the present invention; and

[0056] Fig. 40 is a flow chart showing a procedure of the HTML table reformation processing according to the seventeenth embodiment.

DESCRIPTION OF THE REFERRED EMBODIMENTS

[0057] The present invention will now be explained in connection with the preferred embodiments thereof with reference to the accompanying drawings.

[First Embodiment]

[0058] Fig. 1 is a block diagram showing a functional construction of a document segmentation apparatus according to a first embodiment of the present invention. In Fig. 1, an HTML table storage part 101 serves to hold or store a table (a portion between the tags <table> and </table>) in the HTML document to be processed.

[0059] A table analysis part 102 serves to analyze the table stored in the HTML table storage part 101 and to generate cell position data representing a positional relationship between cells and cell vectors representing characteristics of the cells.

[0060] The cell vector is determined by height and width of the cell, a displaying position of contents, a background color, length and character type of a text in the cell, and magnitude and shape of an image in the cell. The dimension of the cell is (the number of images in the cell x 4 + 17)~~ dimensions~~, and each component is a real number greater than 0 and smaller than 1. When it is assumed that the image which firstly appears in the cell is image$_1$, the k-th component v(k) of the cell vector v is defined as follows:

v(0) : when [[a kind of a]] the tag is <TH> (cell representing item name), 1,0, and when <TD> (cell representing data), 0.0

v(1) : when rowspan (row width) is below 4, rowspan x 0.25, and when the rowspan is above 4, 1.0

v(2) : when ~~clospan~~ colspan (column width) is below 4, colspan x 0.25, and when the ~~clospan~~ colspan is above 4, 1.0

v(3) : when nowrap (no line space) is designated, 1.0, and when not designated, 0.0

v(4) : when align (lateral position) is not designated, 0.0, and when left (left end), 0.2, and when center (central position), 0.4, and when right (right end), 0.6, and when justify (uniform), 0.8, and when others, 1.0

v(5) : when valign (vertical position) is not designated, 0.0, and when top (upper end), 0.2, and when middle (center), 0.4, when bottom (lower end), 0.6, and when baseline, 0.8, and when others, 1.0

v(6) : when bgcolor (background color) is not designated, 0.0, and when not designated by 16-scale code, 0.0, and when designated by 16-scale code, bgcolor/OxFFFFF

v(7) : before ninth row, (row number) x 0.1, and after tenth row, 1.0

v(8) : before 99-th column, (column number) x 0.01, and after 100-th column, 1.0

v(9) : when the number of line spaces (<BR>) is below 5, (<BR> number) x 0.2, and when <BR> number is above 5, 1.0

v(10) : when the number of characters in text is below 100, (number of characters) x 0.01, and when above 100, 1.0

v(11) : (number of numerals in text)/(total number of characters in text)

v(12) : (number of alphabetics in text)/(total number of characters in text) v(13) : (number of Kanji in text)/(total number of characters in text)

v(14) : (number of Katakana in text)/(total number of characters in text)

v(15) : (number of Hiragana in text)/(total number of characters in text)

v(16) : when there is punctuation point "。" or "."). 1.0, and when no punctuation point, 0.0

v(13+ix4) : when an area of image$_1$ is below 150000, (area)/150000, and when above 150000, 1.0

v(14+ix4) : when a height of image$_1$ is below 300, (height)/300, and when above 300, 1.0

v(15+ix4) : when a width of image$_1$ is below 500, (width)/500, and when above 500, 1.0

v(16+ix4) : among character rows representing URL of page containing this table, a ratio of partial character rows common to URL of image$_1$. For example, if an image "../image/hoge.gif"" is included in a page "http://hogehoge.aaa.bbbbb.co.jp: 8080/hogel/hoge2/hoge.html (length of URL is 58), when the image is rewritten to fullpass URL, since "http://hogehoge.aaa.bbbbb.co.jp:8080/hogel/image/ hoge.gif"" is obtained, the common character row becomes "http://hogehoge.aaa.bbbbb.cc.jp:8080/ hogel/". Since this length is 43, a value of this component becomes 43 ÷ 58 = 0, i.e., 741.

[0061] A cell vector storage part 103 is a cell position data storage part for storing cell position data generated by the table analysis part 102. A cell vector storage part 104 serves to store the cell vectors generated by the table analysis part 102.

[0062] A table type ~~judgement~~ judgment part 105 serves to judge a type of the table with reference to the cell position data stored in the cell position data storage part 103 and the cell vectors stored in the cell vector storage part 104 and to instruct a cut direction determination part 107 or a cell cluster generation part 111 to start the processing in dependence upon the table type. There are seven table types from table 1 to table VII which will be described below.

[0063] table I : heights and widths of all of the cells are 1, and the cells in first column/n-th row and n-th column/first row are all <TH> or same background color.

[0064] table II : heights and widths of all of the cells are 1, and the cells in first column/n-th row and n-th column/first row (except for first column/first row) are all <TH> or same background color.

[0065] table III : heights and widths of all of the cells are 1, and the cells in first column/n-th row are all <TH> or same background color.

[0066] table IV : heights and widths of all of the cells are 1, and the cells in first column/n-th row (except for first column/first row) are all <TH> or same background color.

[0067] table V : heights and widths of all of the cells are 1, and the cells in n-th column/first row are all <TH> or same background color.

[0068] table VI : heights and widths of all of the cells are 1, and the cells in n-th column/first row (except for first column/first row) are all <TH> or same background color.

[0069] table VII : tables other than table I to table VI.

[0070] In the above, the tables I to VI are tables showing tables as they are and the table VII is a table used as a tool for the purpose of layout. When the table type is any one of the tables I to VI, the cut direction determination part 107 is instructed to start the processing, and when the table type is the table VII, the cell cluster generation part is instructed to start the processing.

[0071] A table type storage part 106 serves to store the table type determined by the table type ~~judgement~~ judgment part 105.

[0072] When the cut direction determination part 107 is instructed to start the processing by the table type ~~judgement~~ judgment part 105, the part 107 judges whether each data is expressed by column or row in the table describing "table", with reference to the cell position data stored in the cell position data storage part 103 and the cell vectors stored in the cell vector storage part 104, thereby determining the table division direction.

[0073] A score $S_h(T)$ when a table T of N-th column/M-th row is divided on the basis of column and a score $S_v(T)$ when the table T of N-th column/M-th row is divided on the basis of row are defined as follows. In the following description, $\cos(v_{i,j}, v_{k,1})$ represents a cosine value between a table cell vector $v_{i,j}$ in i-th column/j-th row and a table cell vector $v_{k,1}$ in k-th column/l-th row.

[0074] However, these are values calculated only when there are both the data of cell in the i-th column/j-th row and data of cell in the k-th column/l-th row, and if either of both data is not existed, the value becomes zero.

exist(i,j) = 1 (data is existed in the cell in the i-th column/j-th row)
0 (data is not existed in the cell in the i-th column/j-th row)

$$\text{count}_h = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=j+1}^{M} \text{exist}(i, j) \times \text{exist}(i, k)$$

$$\text{count}_v = \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{l=j+1}^{N} \text{exist}(i, j) \times \text{exist}(l, j)$$

$$S_h T = \frac{1}{count} h \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=j+1}^{M} \cos(v_{i,j}, v_{i,k})$$

$$S_v T = \frac{1}{count_V} \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{l=j=1}^{N} \cos(v_{i,j}, v_{l,i})$$

[0075] Since the dimension of the table cell vectors is determined by the number of images includes in the cells in i-th column/j-th row and k-th column/1-th row, the cosine value is calculated by adding component having a value of zero to the lower table cell vectors so that the dimensions of both vectors becomes the same.

[0076] $S_h(T)$ is an average cosine value between two cell table cell vectors in the same column, and $S_v(T)$ is an average cosine value between two cell table cell vectors in the same row. Since the cosine values of the two table cell vectors can be regarded as similarity of cells, it is said that $S_h(T)$ is average similarity between the cells in the same segment when the table [[id]] is divided from column to column and $S_v(T)$ is average similarity between the cells in the same segment when the table [[id]] is divided from row to row.

[0077] Since it is better that the similarity between the cells in the same segment is low in order to incorporate various data in the cells, it is judged that when $S_h(T) \leq S_v(T)$ the table T should be divided from column to column and when $S_h(T) > S_v(T)$ the table T should be divided from row to row.

[0078] A cut direction storage part 108 serves to store the cut directions determined by the cut direction determination part 107.

[0079] A table segment generation part 109 serves to generate the segment from the table describing the table with reference to the table types stored in the table type storage part 106 and the cut directions stored in the cut direction storage part 108. When the cut direction is column direction, in the table of table V type, the columns are made to segments as they are, and, in the tables other than the table V type, the segment is formed by combining the first column. When the cut direction is row direction, in the table of table III type, the rows are made to segments as they are, and, in the tables other than the table III type, the segment is formed by combining the first row.

[0080] A table segment storage part 110 serves to store the table segment generated by the table segment generation part 109.

[0081] A cell cluster storage part 111 serves to effect clustering of cells in the table having the purpose of layout with reference to the cell vectors stored in the cell vector storage part 104 when the starting of processing is instructed by the table type determination part 105. Here, sorting of cells is determined by using maximum distance algorithm. Now, the clustering procedure of the maximum distance algorithm will be described.

[0082] Step 1 : From N (number) sample pattern concurrent X (= x_1, x_2, ..., X_N), any one of sample (for example, x_1) is selected and is made as cluster center $z_1 \in Z$.

[0083] Step 2 : Regarding all of $x_1 \in X$ not included in Z, among cluster centers $z_1 \in Z$, a distance dx_j to the nearest cluster center is calculated. It is assumed that x_i giving Max{dx_1) is x_0.

[0084] Step 3 : Regarding all of $z_k \in Z$, among cluster centers other than z_k, a distance dz_k to the furthest cluster center is calculated.

[0085] Step 4 : When $dx_c \geq \max\{dz_k\} \times t$ ($t = 0.5$ to 1) is established, it is regarded as a new cluster center, and the algorithm is returned to [[the]] Step 2 to select next cluster center. If $dx_c < \max\{dz_k\} \times t$ ($t = 0.5$ to 1), the algorithm goes to Step 5.

[0086] Step 5 : All of $x_1 \in X$ is stored to clusters of the nearest $z_1 \in Z$.

[0087] An example of a clustering result based on the maximum distance algorithm is shown in Fig. 4.

[0088] A cell cluster information storage part 112 serves to store cell cluster information generated by the cell cluster generation part 111.

[0089] A layout segment generation part 113 serves to generate the segments from the table having purpose of layout with reference to the cell position data stored in the cell position data storage page 103 and the cell cluster information stored in the cell cluster information storage part 112.

[0090] The merit ~~for~~ of arranging the information by utilizing the types of the tables is that longitudinal and lateral repeating of a certain arrangement pattern can easily be attained. Thus, the arrangement pattern is guessed on the basis of the cell cluster information, and the segment is obtained by combining the cells matched to the pattern, because, when a certain arrangement pattern appears repeatedly, it can be judged that the cells matched to said pattern are resembled meaningly. Details of processing will now be described.

[0091] First of all, a fundamental cell kind is determined, and cells in the fundamental cell kind are regarded as fundamental cells. The fundamental cell kind is selected as a cell kind having least number of cells among cell kinds including a plurality of same cells. If there are a plurality of cell kinds in question, the leftmost or uppermost cell kind is selected.

[0092] Then, it is ascertained whether any cell having the same kind of the cell adjacent to the fundamental cell is adjacent to another fundamental cell or not. If adjacent, the fundamental cells are connected to obtain a new fundamental cell. This procedure is repeated until the cells cannot be interconnected.

[0093] When the above-mentioned process is finished, the fundamental cell and the remaining cells are made to segments, respectively.

[0094] A layout segment storage part 114 serves to store the layout segments generated by the layout segment generation part 113. The table segments stored in the table segment storage part 110 and the layout segments stored in the layout segment storage part 114 are segments eventually obtained.

[0095] Fig. 2 is a view showing a hardware arrangement of the document segmentation apparatus according to the illustrated embodiment.

[0096] In Fig. 2, a CPU 201 serves to effect the processing in accordance with program stored in a ROM 202. The ROM 202 serves to store program performing control procedure which will be described later. A RAM 203 serves to provide storing areas required for operating the cell position data storage part 103, cell vector storage part 104, table type storage part 106, cut direction storage part 108, cell cluster information storage part 112 and the aforementioned program.

[0097] A disk drive device 204 serves to realize the HTML table storage part 101, table segment storage part 110, and layout segment storage part 114. A bus[[s]] 205 serves to connect between the above-mentioned elements and permit sending and receiving of data between the elements.

[0098] Next, a processing operation of the illustrated embodiment will be explained. Fig. 3 is a flow chart showing an operation procedure of the document segmentation apparatus according to the illustrated embodiment.

[0099] In [[a]] step S301, the tables stored in the HTML table storage part 101 are analyzed to generate the cell position data representing the positional relationship between the cells and the cell vectors representing characteristics of the cells. Then, the program goes to [[a]] step S302.

[0100] In [[the]] step S302, the table type is determined with reference to the cell position data stored in the cell position data storage part 103 and the cell vectors stored in the cell vector storage part 104. Then, the program goes to [[a]] step S302.

[0101] In [[the]] step S303, it is judged whether the table to be processed is the table describing the table or not with reference to the table types stored in the table type storage part 106. If the table is the table describing the table, the program goes to [[a]] step S304. If not, the program goes to [[a]] step S306.

[0102] In [[the]] step S304, it is determined whether the data in the table describing the table are represented by column or row with reference to the cell position data stored in the cell position data storage part 103 and the cell vectors stored in the cell vector storage part 104, thereby determining the dividing direction of the table. Then, the program goes to [[a]] step S305.

[0103] In [[the]] step S305, the segments are generated from the table showing the table as it is with reference to the table types stored in the table type storage part 106 and the cut directions stored in the cut direction storage part 108, and ~~. And,~~ the operation is finished.

[0104] In [[the]] step S306, the cells in the table used as a tool for the purpose of layout are clustered with reference to the cell vectors stored in the cell vector storage part 104. Then, the program goes to [[a]] step S307.

[0105] In [[the]] step S307, the segments are generated from the table describing the table with reference to the cell position data stored in the cell position data storage part 103 and the cell cluster information stored in the cell cluster information storage part 112, and ~~. And,~~ the operation is finished.

[0106] As mentioned above, by analyzing the table to be processed to judge whether the table is the table describing the table or the table having purpose of layout and by generating the segments by effecting the processing for obtaining the target table, the tables in the HTML document can be divided according to their contents.

[Alterations]

[0107] In the above-mentioned embodiments, while an example that the maximum distance algorithm is used for effecting the clustering of the cells was explained, the present invention is not limited to ~~such an~~ that example, but[[,]] the clustering may be effected by using other algorithms.

[0108] The definition of the components of the cell vectors shown in the illustrated embodiment is merely one example, and[[,]] the characteristics of the cells may be expressed by other definitions.

[0109] The definition of score for determining the cut direction shown in the illustrated embodiment is merely one example, and[[,]] the cut direction may be determined by other definitions.

[0110] In the illustrated embodiment, while an example that the height and width of the cell, kind of the tag (TH or TD) and the background color are used to determine the column (or row) of the item name for determining the table type was explained, the present invention is not limited to such an example, but[[,]] ~~judgement~~ judgment may be effected by using other attributes.

[0111] In the illustrated embodiment, while an example that the cell position data storage part 103, cell vector storage part 104, table type storage part 106, cut direction storage part 108 and cell cluster information storage part 112 are realized by the RAM and the HTML table storage part 101, table segment storage part 110 and layout segment storage part 114 are realized by the disk drive device was explained, the present invention is not limited to such an example, but[[,]] these may be realized by using any recording medium.

[0112] In the illustrated embodiment, while an example that the HTML table is divided was explained, so long as the contents of the table can be discriminated, another type of table may be divided.

[0113] In the illustrated embodiment, while an example that the elements are incorporated into the same computer was explained, the present invention is not limited to such an example, but[[,]] the elements may be individually incorporated into computers or processing devices included in a network.

[0114] In the illustrated embodiment, while an example that the program is stored in the ROM was explained, the present invention is not limited to such an example, but, the program may be stored in any recording medium. Further, the program may be realized by any circuit performing the same operation.

[Second Embodiment]

[0115] In the above-mentioned embodiment, while an example ~~that~~ in which the apparatus serves to divide only ~~the~~ HTML tables was explained, the present invention is not limited to ~~such an~~ that example. For example, the present invention may be realized as an apparatus for dividing the entire HTML document. Fig. 5 is a block diagram showing [[a]] the fundamental construction in an example of such a case.

[0116] In Fig. 5, an HTML document storage part 501 serves to store an HTML document to be processed. A normal segment generation part 502 serves to divide the HTML document stored in the HTML document storage part 501 into segments. A normal segment storage part 503 serves to store segments other than tables generated by the normal segment generation part 502. An HTML table storage part 101 serves to store segments of tables generated by the normal segment generation part 502. ~~Others~~ Other parts are the same as those shown in Fig. 1.

[0117] In Fig. 5, the normal segments stored in the normal segment storage part 503, the table segments stored in the table segment storage part 110 and the layout segments stored in the layout segment storage part 114 are the segments eventually obtained.

[Third Embodiment]

[0118] In the above-mentioned embodiments, while an example ~~that~~ in which both ~~the~~ table tags actually showing ~~the~~ a table ~~as it is~~ and ~~the~~ table tags used as a tool for the purpose of layout are divided into ~~the~~ segments was explained, the present invention is not limited to ~~such an~~ that example. For example, only ~~the~~ a table ~~showing the~~ that is an actual table ~~as it is~~ may be divided. Fig. 6 is a block diagram showing [[a]] the fundamental construction in an example of such a case.

[0119] In Fig. 6, when a table segment generation part 601 is instructed to start processing from a table type ~~judgement~~ judgment part 105, it generates an HTML table stored in an HTML table storage part 101 as table segments.

[0120] A table segment storage part 602 serves to store table segments generated by a table segment generation part 611. ~~Others~~ Other parts are the same as those shown in Fig. 1.

[0121] In Fig. 6, the table segments stored in the table segment storage part 110 and the table segments stored in the table segment storage part 602 are the segments eventually obtained.

[Fourth Embodiment]

[0122] In the above-mentioned embodiments, while an example ~~that~~ in which both ~~the~~ table formatting that is used to show~~ing the~~ an actual table ~~as it is~~ and ~~the~~ table formatting used as a tool for the purpose of layout are divided into ~~the~~ segments was explained, only the table used as a tool for the purpose of layout may be divided. Fig. 7 is a block diagram showing [[a]] the fundamental construction in an example of such a case.

[0123] In Fig. 7, when a table segment generation part 701 is instructed to start processing from a table type ~~judgement~~ judgment part 705, it generates an HTML table stored in an HTML table storage part 101 as table segments. A table segment storage part 702 serves to store table segments generated by a table segment generated part 706. ~~Others~~ Other parts are the same as those shown in Fig. 1.

[0124] In Fig. 7, the table segments stored in the table segment storage part 702 and the layout segments stored in the layout segment storage part 114 are the segments eventually obtained.

[0125] Incidentally, in the above-mentioned embodiment, while an example ~~that~~ in which the present invention is applied to ~~the~~ apparatus for dividing the HTML document was explained, the present invention is not limited to ~~such an~~ that example, but[[,]] the present invention may be realized as a segment retrieving apparatus in which retrieval can be effected for each segment unit by combining the dividing apparatus with a retrieving apparatus.

[Fifth Embodiment]

[0126] In the above-mentioned embodiments, while an example ~~that~~ in which the ~~judgement~~ judgment as to whether the table formatting is ~~the~~ for providing an actual table ~~showing the table as it is~~ or not is effected only on the basis of the syntax of the table, was explained.

[0127] However, among the HTML documents tables, since there are also tables in which table items are not described by emphasizing characters to permit discrimination as TH tags or item name, ~~the~~ it is possible that a table formatting that is being used for describing ~~the~~ an actual table may be erroneously judged as being for layout. In such a case, the approach based only ~~from~~ on the syntax has a limitation ~~for~~ as to judging whether the table ~~is the table describing the~~ represents an actual table or not.

[0128] Now, referring to an example shown in Fig. 8, since meanings between the cells are analogous with each other, it can be seen that each cell forms an element for one item. In this way, among the HTML document tables, there are also tables which can be discriminated as ~~table~~ ones showing actual tables ~~as it is~~ by semantics.

[0129] Thus, in a fifth embodiment of the present invention, the ~~judgement~~ judgment as to whether ~~the~~ a table ~~is the table describing the~~ shows an actual table or not is effected on the basis of ~~approach from the~~ semantics.

[0130] Fig. 9 is a block diagram showing a construction of an apparatus according to the fifth embodiment.

[0131] In a table analysis part 102, a table stored in an HTML table storage part 101 is analyzed to generate cell position data representing a positional relationship between cells, cell vectors representing characteristics of the cells and data for cells. A cell data storage part 901 serves to store the cell data generated by the table analysis part 102. ~~Others~~ Other parts are the same as those shown in Fig. 1.

[0132] The processing procedure according to the illustrated embodiment is effected in accordance with the flow chart shown in Fig. 3, as [[is]] in the first embodiment. However, since there [[is]] are slight ~~detailed~~ differences in detail from the first embodiment, such differences will be described.

[0133] In [[a]] step 301, the table stored in the HTML table storage part 101 is analyzed to generate cell position data representing a positional relationship between cells, cell vectors representing characteristics of the cells and data for cells. ~~And,~~ Then the program goes to [[a]] step S302.

[0134] In [[the]] step S302, the table type is determined with reference either the cell position data stored in the cell position storage page 103 or the cell vectors stored in the cell vector storage part 104 or the cell data stored in the cell data storage page 901, and ~~. And,~~ the program goes to [[a]] step S303.

[0135] Here, the determination of the table type includes determination of table type on the basis of a thesaurus, determination of table type on the basis of similarity of character, determination of table type on the basis of syntax and determination of table type on the basis of coincidence of character. An operation for determining the table type will be described in connection with embodiments which will be described ~~later~~ below. ~~The step~~ Step S303 and other steps are the same as those in the first embodiment.

[0136] In the illustrated embodiment, the table ~~judgement~~ judgment part 105 includes a thesaurus similarity ~~judgement~~ judgment part 1001 and a thesaurus dictionary 1002. Now, [[an]] its operation will be explained with reference to Fig. 10.

[0137] The term "thesaurus" is a word ~~for~~ meaning a high/low rank relationship between vocabularies. Words include high rank words, which are more abstract, a synonym for a given word (no ~~which does not~~ change in meaning even if expressed by the other word), analogous words, which are ~~resembled~~ similar in meaning, and low rank words, which are more concrete. For example, a word "morning glory" includes "flower" as the high rank word and "violet", "convolvulus" and "balsam" as analogous words. [[A]] The word "flower" includes "violet", "convolvulus" and "balsam" as ~~the~~ low rank words.

[0138] The thesaurus similarity ~~judgement~~ judgment part 1001 serves to judge the table type on the basis of thesaurus similarity described in the thesaurus dictionary 1002 with reference to the cell position data stored in the cell position data storage part 103 and the cell data stored in the cell data storage part 115, and the judged table type is stored in the table type storage part 106.

[0139] Now, the ~~judgement~~ judgment of the table type based on the thesaurus similarity will be explained with reference to an example of an M column/N row table.

[0140] A function for obtaining score based on the thesaurus for two character lines sl, s2 is expressed as f(sl, s2). When the character line s2 is the synonym or analogous word with the character line sl, the value of f(sl, s2) becomes maximum. It is assumed that, as the character line s2 with respect to the character line sl becomes gradually deeper in the high rank word direction or the low ~~rand~~ rank word direction, the value f(sl, s2) becomes smaller.

[0141] When it is assumed that a character line of m-th column/n-th row cells is $S_{m,n}$, the average score of thesaurus for cells in the first row can be expressed as follows:

$$\frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M} f\left(S_{i,1}, S_{j,1}\right)$$

[0142] Similarly, the average score of thesaurus for cells in the first column can be expressed as follows:

$$\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N} f\left(S_{i,1} S_{1,j}\right)$$

[0143] If the average score of thesaurus for cells in the first column or row exceeds a threshold value, it is judged ~~as the~~ that the table formatting is being used for describing ~~the~~ an actual table, and, if the average score does not exceed the threshold value, it is judged [[as]] that the table formatting is being used for describing the layout. In this way, the table type of the table to be processed can be judged.

[0144] As a method for obtaining the score based on similarity of character regarding two character lines sl, s2, there is a method called [[as]] “vague retrieval”.

[0145] A function for obtaining score based on the ~~similarly~~ similarity of character for two character lines sl, s2 is expressed as g(sl, s2). When it is assumed that if the similarity of character is great, the [[a]] value of g(sl, s2) becomes greater and if the similarity of character is ~~great~~ small, the value of g(sl, s2) becomes smaller, by using the vague retrieval, similar to the method for obtaining the score on the basis of the thesaurus:[[,]] if the average score of similarity of character for cells in the first column or row exceeds a threshold value, it is judged as ~~the~~ being an actual ~~table describing the~~ table, and[[,]] if the average score does not exceed the threshold value, it is judged [[as]] that the table formatting is being used for describing the layout. In this way, the table type of the table to be processed can be judged.

[0146] In the illustrated embodiment, regarding the table to be processed, first of all, the ~~judgement~~ judgment of the table based on the thesaurus is effected, and, if the table is ~~the table describing the~~ an actual table, the procedure is ended, ~~and~~ while, if the table is not ~~the table describing the~~ an actual table, the table ~~judgement~~ judgment based on the similarity of character is effected regarding the table to be processed.

[0147] In this way, the table type of the table to be processed can be effected on the basis of the thesaurus similarity.

[0148] Now, the details of the table ~~judgement~~ judgment in [[the]] step S302 will be explained with reference to Fig. 11.

[0149] In [[a]] step S1101, from the cell position data stored in the cell position data storage part 103 and the cell data stored in the cell data storage part 901, the type of the table to be processed is judged on the basis of the thesaurus, and, if the table is ~~the table describing the~~ an actual table, the procedure is ended, ~~and~~ while, if the table is not ~~the table describing the~~ an actual table, the program goes to [[a]] step S1102.

[0150] In [[the]] step S1102, from the cell position data and the cell data, the type of the table to be processed is judged on the basis of the similarity of character. Then, the procedure is ended.

[0151] Here, an example of the table of a page regarding "How to Rear Flowers" shown in Fig. 8 will be explained. First of all, the average scores of thesaurus for cells in the first column and the first row are measured. In the first column, it can be seen that words "violet", "morning glory" and "balsam" are included. These words are words ~~regarding the~~ having to do with flowers. Accordingly, the average score of thesaurus regarding the cells in the first column becomes great, and, thus, this table can be judged as ~~the table describing the~~ being an actual table.

[0152] Next, an example of a table regarding "A Page of Products Catalog" shown in Fig. 12 will be explained. First of all, the average scores of similarity of character for cells in the first column and the first row are measured. In the first column, it can be seen that words "AAA0001", "AAA0002" and "AAA1001" are included. These words are analogous words. Accordingly, the average score of similarity of character regarding the cells in the first column becomes great, and, thus, this table can be judged as ~~the table describing the~~ being an actual table.

[0153] As mentioned above, by analyzing the table to be processed on the basis of the semantics to judge whether the table is ~~the table~~ describing the table or ~~the table~~ one having the purpose of layout and by generating the segments accordingly, the table in the HTML document can be divided ~~from content~~ to produce segments differing from each other in content.

[Sixth Embodiment]

[0154] In a sixth embodiment of the present invention, a table ~~judgement~~ judgment portion 105 includes a partial character line extracting part 1301 and a character line comparison part 1302. [[An]] Its operation will be explained with reference to Fig. 13.

[0155] In the partial character line extracting part 1301, partial character lines of the cells are extracted with reference to the cell position data stored in the cell position data storage part 103 and the cell data stored in the cell data storage part 901. The extraction of the partial character line is effected by using a known method, such as geometric element analysis.

[0156] In the character line comparison part 1302, the partial character lines of the cells extracted in the partial character line extracting part 1301 are compared, so that the table type is judged depending upon whether the character lines ~~are~~ coincide[[d]] with each other in many cells or not. The judged table type is stored in the table type storage part 106.

[0157] Now, the ~~judgement~~ judgment of the table type based on the character line comparison will be explained with reference to an example of an M-th column/N-th row table.

[0158] A function for obtaining coincidence of character line regarding two character lines sl, s2 is expressed as h(sl, s2). It is assumed that, if h(sl, s2) ≠ 0, the two character lines do not coincide, and, if h(sl, s2) = 0, the two character lines do coincide with each other.

[0159] When it is assumed that a character line of m-th column/n-th row cell is $S_{m,n}$ and a k-th partial character line from the top when $S_{m,n}$ is divided into the partial character lines is $S^k_{m,n}$, an average of coincidence of character line regarding the last character lines of the cells in the first column can be expressed as follows:

$$\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M}\left|h\left(S^{m}_{i,1},S^{n}_{j,1}\right)\right|$$

[0160] $S^m_{i,j}$, $S^n_{j,i}$ represent the last partial character lines in the respective character lines. Similarly, an average of coincidence of character line regarding the last character lines of the cells in the first row can be expressed as follows:

$$\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left|h\left(S^{m}_{1,i},S^{n}_{1,j}\right)\right|$$

[0161] If the average of coincidence of character line regarding the cells in the first column or the first row does not exceed a threshold value, it is judged as ~~the table describing the~~ being an actual table, and, if the average exceeds the threshold value, it is judged as the table describing the layout. In this way, the type of the table to be processed can be judged. After the processing, the judged table type is stored in the table type storage part 106. In this way, the table type can be judged on the basis of the character line comparison.

[0162] Now, the details of the table judgement in [[the]] step S302 will be explained with reference to Fig. 14.

[0163] In [[a]] step S1401, from the cell position data and the cell vectors, the partial character line is extracted. And, the program goes to [[a]] step S1402.

[0164] In [[the]] step S1402, the partial character lines of the cells are compared, and the table type is judged depending upon whether the character lines are coincided with each other in many cells or not. And the procedure is ended.

[0165] Now, an example of a table regarding "A Page of Medical Centers" shown in Fig. 15 will be explained.

[0166] First of all, the cells in the first column and the first row are divided into partial character lines by using the geometric element analysis. When the cells in the first row are divided into the partial character lines, "A clinic", "B clinic" and "C clinic" are obtained. When the character line comparison is effected between the last partial character lines of the cells, since "clinic" coincides, the average of coincidence of character line regarding the cells in the first row becomes small, and, thus, it can be judged as the table describing the table.

[0167] As mentioned above, by analyzing the coincidence of partial character line of cells to judge whether the table to be processed is the table showing the table or the table having purpose of layout and by generating the segments accordingly, the table in the HTML document can be divided from content to content.

[Seventh Embodiment]

[0168] In a seventh embodiment of the present invention, a table judgement portion 105 includes a partial character line extracting part 1601, thesaurus similarity judgement part 1602, and a thesaurus dictionary 1603. An operation will be described with reference to Fig. 16.

[0169] In the partial character line extracting part 1301, a partial character lines are extracted with reference to the cell position data stored in the cell position data storage part 103 and the cell data stored in the cell data storage part 115.

[0170] In the thesaurus similarity judgement part 1602, regarding the partial character lines of the cells extracted in the partial character line extracting part 1301, the table type is judged on the basis of thesaurus similarity of the thesaurus dictionary 1603, and the judged table type is stored in the table type storage part 106.

[0171] Now, the details of the table judgement in [[the]] step S302 will be explained with reference to Fig. 17.

[0172] In [[a]] step S1701, from the cell position data and the cell vectors, the partial character line is extracted. And, the program goes to [[a]] step S1702.

[0173] In [[the]] step S1702, regarding the partial character liens of the cells, the table judgement based on thesaurus is effected. As a result, in [[a]] step 1703, if the table is the table describing the table, the procedure is ended; otherwise, the program goes to [[a]] step S1704.

[0174] In [[the]] step S1704, regarding the partial character lines of the cells, the table judgement based on similarity of character is effected. And, the procedure is ended.

[0175] As mentioned above, by judging the table type of the table to be processed on the basis of the thesaurus similarity regarding the partial character lines of the cells to judge whether the table is the table showing the table or the table having purpose of layout and by generating the segments accordingly, the table in the HTML document can be divided from content to content.

[Eighth Embodiment]

[0176] In an eighth embodiment of the present invention, a table judgement portion 105 includes a syntax judgement part 1801, a thesaurus similarity judgement part 1802, and a thesaurus dictionary 1803. An operation will be described with reference to Fig. 18.

[0177] The syntax judgement part 1801 serves to effect the processing similar to the table type judgement part 105 of the first embodiment. After the processing in the syntax judgement part 1801 or the thesaurus similarity judgement part 1802, the judged table type is stored in the table type storage part 106.

[0178] Now, the details of the table judgement in [[the]] step S302 will be explained with reference to Fig. 19.

[0179] In [[a]] step S1901, from the cell position data and the cell vectors, the table type is judged on the basis of syntax. As a result, in [[a]] step 1902, if the table is the table describing the table, the procedure is ended; otherwise, the program goes to [[a]] step S1903.

[0180] In [[the]] step S1903, from the cell position data and the cell vectors, the table type is judged on the basis of thesaurus. As a result, in [[a]] step 1904, if the table is the table describing the table, the procedure is ended; otherwise, the program goes to [[a]] step S1905.

[0181] In [[the]] step S1905, from the cell position data and the cell vectors, the table type is judged on the basis of similarity of character. And, the procedure is ended.

[0182] As mentioned above, by analyzing the table type of the table to be processed on the basis of syntax and semantics to judge whether the table is the table showing the table or the table having purpose of layout and by generating the segments accordingly, the table in the HTML document can be divided from content to content.

[Ninth Embodiment]

[0183] In a ninth embodiment of the present invention, a table ~~judgement~~ judgment portion 105 includes a syntax ~~judgement~~ judgment part 2001, a partial character line extracting part 2002, and a character line comparison part 2003. An operation will be described with reference to Fig. 20.

[0184] The syntax ~~judgement~~ judgment part 1801 serves to effect the processing similar to the table type ~~judgement~~ judgment part 105 of the first embodiment. The partial character line extracting part 2002 and the character line comparison part 2003 serve to effect the processing similar to the partial character line extracting part 1301 and the character line comparison part 1302 of the sixth embodiment. After the processing in the syntax ~~judgement~~ judgment part 2001 or the character line comparison part 2003, the judged table type is stored in the table type storage part 106.

[0185] Now, the details of the table ~~judgement~~ judgment in [[the]] step S302 will be explained with reference to Fig. 21.

[0186] In [[a]] step S2101, from the cell position data and the cell vectors, the table type is judged on the basis of syntax. As a result, if the table is the table describing the table, the procedure is ended; otherwise, the program goes to [[a]] step S2102.

[0187] In [[the]] step S2102, from the cell position data and the cell vectors, the partial character lines are extracted, and, in [[a]] step S2103, the partial character lines of the cells are compared, so that the table type is judged depending upon whether the partial character lines ~~are~~ coincide[[d]] with each other in many cells or not. ~~And,~~ Then, the procedure is ended.

[0188] As mentioned above, by analyzing the table type of the table to be processed on the basis of syntax and the coincidence of the partial character line to judge whether the table is the table describing the table or the table having purpose of layout and by generating the segments accordingly, the table in the HTML document can be divided from content to content.

[Tenth Embodiment]

[0189] In a tenth embodiment of the present invention, a table judgement portion 105 includes a syntax judgement part 2201, a partial character line extracting part 2202, a thesaurus similarity judgement part 2203 and a thesaurus dictionary. An operation will be described with reference to Fig. 22.

[0190] The syntax judgement part 2201 serves to effect the processing similar to the table type judgement part 105 of the first embodiment. The partial character line extracting part 2202 and the thesaurus similarity judgement part 2203 serve to effect the processing similar to the partial character line extracting part 1601 and the thesaurus similarity judgement part 1602. After the processing in the syntax judgement part or the thesaurus similarity judgement part, the judged table type is stored in the table type storage part 106.

[0191] Now, the details of the table judgement in [[the]] step S302 will be explained with reference to Fig. 23.

[0192] In [[a]] step S2301, from the cell position data and the cell vectors, the table type is judged on the basis of syntax. As a result, in [[a]] step S2302, if the table is the table describing the table, the procedure is ended; otherwise, the program goes to [[a]] step S2303.

[0193] In [[the]] step S2303, from the cell position data and the cell vectors, the partial character lines are extracted, and, in [[a]] step S2304, regarding the partial character lines of the cells, the table judgement is effected on the basis of thesaurus. As a result, in [[a]] step S2305, if the table is the table describing the table, the procedure is ended; otherwise, the program goes to [[a]] step S2306. In [[the]] step S2306, regarding the partial character ~~liens~~ lines of the cells, the table judgement is effected on the basis of similarity of character. And, the procedure is ended.

[0194] As mentioned above, by analyzing the table type of the table to be processed on the basis of syntax and analyzing the partial character lines of the cells to judge whether the table is the table describing the table or the table having purpose of layout and by generating the segments accordingly, the table in the HTML document can be divided from content to content.

[0195] In the above-mentioned embodiments, when the judgement whether the table is the table describing the table or not, by utilizing the table judgement based on semantics as well as the table judgement based on syntax, regarding many tables, it is possible to judge whether such table is the table describing the table or not.

[Eleventh Embodiment]

[0196] Now, naming regarding the table will be briefly described.

[0197] "Record" is information representing one substance, and a group of records representing similar substances constitute record concurrence. Of course, styles of the records in the record concurrence are the same. The record is constituted by fields (data) representing attributes of the substances. For example, "Taro

Yamada: Yokohama-city: 045-000-0000" is a record constituted by three fields. "Hanako Yamada: Kawasaki-chi: 044-111-1111" is also a record representing a person in the same manner as the above record. The concurrence constituted by these two records is recorded concurrence.

[0198] In order to discriminate the fields, since first field, second field and the like are difficult to be understood, naming is frequently used. The naming or title given to the field is called as a field name. For example, in the aforementioned record, it is assumed that the field name of the first field is "(person's) name", second field is "address" and third field is "phone number". Thus, in the first record, a field value of the field name "name" is "Taro Yamada" and a field value of the field name "address" is "Yokohama-city".

[0199] Data actually representing the record concurrence is shown in Fig. 24. In case of the HTML document, the table is concretely described as a table (table is data described by TABLE tags). Fig. 24 shows an example of the record concurrence described by the table.

[0200] In this example, while each column of the table describes one record, there is a case where the rows describe the records. However, since the column and the row may be interchanged, i.e., the column and the row may be converted with respect to a diagonal of the table, in the following explanation, it is regarded that the records are described in the column direction. In the case where the columns represent the records, the readings of column and row are changed, the same result is achieved. In the table shown, the first line describes the fields names of the fields. Such a line is referred to as a field name describing line (i.e., line with the field name). The second and third lines describe one record, respectively. Such a line is referred to as a record describing line (i.e., line with record).

[0201] In the aforementioned embodiments, in order to judge whether the table is the table describing the table is or not, the judgement was effected under a assumption of the table in which M columns and N rows are not omitted and regular description is made. However, among the tables in the HTML document, there are tables in which a plurality of tables are included in one table or the record straddles between plural table. Further, there are also multi-row and multi-column tables in which, when the adjacent informations are the same, the informations are gathered to be described as single information. Regarding such tables, the table judgement cannot be effected easily.

[0202] For these tables, by analyzing a structure of the table and regularity of information description constituting the table, and by reforming the table regularly in M columns and N rows, the table can be divided correctly.

[0203] Fig. 25 is a block diagram showing a construction of an apparatus according to an embodiment of the present invention.

[0204] An HTML table reformation part 2501 serves to reform a table stored in the HTML table storage part 101 regularly without omission of M columns and N rows by analyzing the structure of the table and regularity of information description constituting the table.

[0205] An HTML table reformation data storage part 2502 serves to store data of the HTML table reformed in the HTML table reformation part 2501.

[0206] A table analysis part 102 serves to analyze the table stored in the HTML table reformation data storage art 2502 thereby to generate cell position data indicating a positional relationship between the cells, and cell vectors representing characteristics of the cells and data of the cells. The other constructions are the same as those shown in Fig. 1.

[0207] Next, an operation of the document dividing apparatus according to the illustrated embodiment will be explained with reference to a flow chart shown in Fig. 26.

[0208] In [[a]] step S2600, regarding the table stored in the HTML table storage part 101, by analyzing the structure of the table and regularity of information description constituting the table, the table is reformed regularly without omission of M columns and N rows. And, the program goes to [[a]] step S2601.

[0209] The table reformation includes table reformations based on supplementary data removal, treatment of a multi-row/multi-column table and treatment of a composite table. In the illustrated embodiment, the table reformation is effected by the supplementary data removal. The table reformations based on the treatment of a multi-row/multi-column table and treatment of a composite table will be described in connection with other embodiments. Steps S2601 to S2607 are the same as [[the]] steps S301 to S307 in Fig. 3.

[0210] In the illustrated embodiment, the supplementary data removal is effected by the HTML table reformation part 2501. Here, referring to the table data stored in the HTML table storage part 101, unnecessary data added to the table in the table is removed.

[0211] Next, the details of the HTML table reformation in [[the]] step S2600 will be explained with reference to Fig. 27.

[0212] In [[a]] step S2701, a region of the field name describing line (line with field name) with the TH tags is judged, and in [[a]] step S2702, a region of the field name describing line with tags describing the background color is judged, and, in [[a]] step S2703, a region of the field name describing line with tags for bold face is checked, and the program goes to [[a]] step S2704.

[0213] In [[the]] step S2704, on the basis of the regions of the lines with the field name checked in [[the]] steps S2701 to S2703, meaning similarity between the field names of the lines with the field name and fields perpendicular to the describing directions of the lines with the field name is calculated. Since the field having high score of similarity is description in the field name, by judging the region having high score of similarity, the region in the table is judged. In [[a]] step S2705, the similarity of character line is calculated in the same procedure as [[the]] step S2704 to judge the region in the table.

[0214] In [[a]] step S2706, on the basis of the regions in the table checked in [[the]] steps S2704 to S2705, excessive data other than the table is removed.

[0215] Now, the operation for the supplementary data removal will be described by using a sample. Fig. 28 shows a page of "How to Rear Flowers", in which the supplementary data other than the table are added to the first and fourth columns.

[0216] First of all, in [[the]] steps S2701 to S2703, the lines with the field name are specified. In Fig. 28, since there are field name describing lines with bold face in the second line, by the processing in [[the]] step S2703, it is judged that the second line is the field name describing line.

[0217] Then, in [[the]] steps S2704 and S2705, the region in the table, i.e., a range of the field value regarding the field name is specified on the basis of the similarity of thesaurus or similarity of character line. In Fig. 28, from third to fifth lines in the first column, since "violet", "morning glory" and "balsam" which are the field values regarding the field name "flower name" are described, by the processing in [[the]] step S2704, it is judged that the table has the region corresponding to second to fifth lines.

[0218] Lastly, by the processing in [[the]] step S2706, by removing the supplementary data out of the region in the table, the contents of the table can be picked up.

[0219] As mentioned above, regarding the table to be processed, by analyzing the structure of the table [[an]] and regularity of information description constituting the table and by reforming the table regularly in M columns and N rows, the table can be divided correctly.

[Twelfth Embodiment]

[0220] In a twelfth embodiment of the present invention, the HTML table reformation part 2501 effects the multi-row/multi-column table treatment. Here, by analyzing the structure of the table with reference to the table data stored in the HTML table storage part 101, the table is reformed regularly without omission of M columns and N rows.

[0221] Next, the details of the HTML table reformation in [[the]] step S2600 will be explained with reference to Figs. 29A to 29E.

[0222] When the multi-row/multi-column table is stored every similar tables, (1) by corresponding the structure of the field of the line with the field name to the structure of the field of the record portion, the record can be picked up, (2) the record can be picked up by matching the structure of the field of the field name describing line with the structure of the field of the record, and (3) the record can be picked up by re-reading the field portion including the multi-row/multi-column. A flow of the process regarding (1) is shown in Figs. 29A to 29C, a flow of the process regarding (2) is shown in Fig. 29D and a flow of the process regarding (3) is shown in Fig. 29E.

[0223] When the data in the table including the multi-row/multi-column is handled, the field of the multi-row or multi-column is divided into minimum unit fields which are in turn stored. In this case, regarding the data of the fields of the multi-row/multi-column, the same data are stored in the respective fields at the stage of division. For example, the multi-row/multi-column shown in Fig. 30A is divided into the minimum unit fields which are in turn stored. Thus, as shown in Fig. 30B, a table having four columns and four rows.

[0224] In the above (1), by corresponding the structure of the field of the line with the field name to the structure of the field of the record portion, the record is picked up.

[0225] First of all, a process for analyzing the structure of the field of the line with the field name will be explained with reference to Fig. 29A.

[0226] In [[a]] step S2901, if the field exists, the program goes to [[a]] step S2902. If the field does not exist, the processing of the multi-row/multi-column is ended.

[0227] In [[the]] step S2902, data of a line is extracted, and, in [[a]] step S2903, a region of lines with the field name is judged, and then the program goes to [[a]] step S2904. The region of lines with the field name can be judged by examining different columns in fields in one line presently stored and in the fields in the immediately previous line.

[0228] For example, in the multi-row/multi-column as shown in Fig. 30C, since the data is stored by dividing into the minimum unit fields, as shown in Fig. 30D, the table having four columns and four rows is obtained. Here, when the same data between the fields in the first and second lines is examined, since the fields coincide in the first and fourth columns, the border between the first and second lines is not the border for the line with the field name. However, when the same data between the fields in the second [[nd]] and third lines is examined, since any fields do not coincide, the border between the second and third lines becomes border for the line with the field name. In this way, the structure of the lines with the field name can be grasped.

[0229] In [[the]] step S2904, if the structure of the lines with the field name can be grasped, the program goes to ① . If not grasped, in [[a]] step S2905, data of one line is stored, and, in [[a]] step S2906, it is examined which structures are given in the fields with the field name till the lines which has been examined up to now, and the program is returned to [[the]] step S2901.

[0230] Next, the processing for picking up the records on the basis of the analyzed structures of the fields with the field name will be explained with reference to Fig. 29B. Here, the records in the table in which the structure of the field of the lines with the field name are the same as the structure of the fields of the records, as shown in Fig. 30E, can be picked up. Further, the field is started from the field in the first record.

[0231] In [[a]] step S2907, if the field exists, the program goes to [[a]] step S2908. If the field does not exist, the program goes to [[a]] step S2910. However, if no field exists at all, the processing of the multi-row/multi-column is ended.

[0232] In [[the]] step S2908, the data of one line is extracted, and, in [[a]] step S2909, if the structure of the field of the line with the field name coincides with the structure of one record, the program is returned to [[the]] step S2907. If it does not coincide, the program goes to ②.

[0233] In [[the]] step S2910, on the basis of the structure of the field of the line with the field name, the field information is reformed.

[0234] Next, the processing for picking up the record on the basis of the analyzed structure of the field of the line with the field name will be further explained with reference to Fig. 29C. Here, by the structure of the field having the field value as shown in Fig. 30F, the record can be picked up from the table in which the corresponding field name described lines are different. In this table, the field name describing lines are constituted by a plurality of lines. Thus, regarding the fields in the lines with the field name, by scanning the record coinciding with the structure of the field up to the last line of the table to examine correspondence, the records in the table can be picked up.

[0235] In [[a]] step S2911, if the field name of the field name describing line exists, the program goes to [[a]] step S2912. If it does not exist, the program goes to [[a]] step S2918. However, if no field name exists at all, the processing of the multi-row/multi-column is ended.

[0236] In [[the]] step S2912, the data of one line with the field name is extracted, and, in [[a]] step S2913, if the extracted data of one line does not reach the last line of the lines with the field name, the program goes to [[a]] step S2914. If reached and if data of one line cannot be extracted, the program goes to ③.

[0237] In [[the]] step S2914, if there is a field other than the field of the line with the field name, the program goes to [[a]] step S2915. If it does not exist, the program is returned to [[the]] step S2911. However, if no field exists at all, the processing of the multi-row/multi-column is ended.

[0238] In [[the]] step S2915 the data of one line is extracted, and, in [[a]] step S2916, if the structure of the field of one line with the field name coincides with the structure of the field of one line extracted in [[the]] step S2915, the program goes to [[a]] step S2917. If it does not coincide, the program is returned to [[the]] step S2914.

[0239] In [[the]] step S2917, the structure information of the field name describing line to which the line presently scanned coincides is stored, and the program is returned to [[the]] step S2914.

[0240] In [[the]] step S2918, on the basis of the structure information stored in [[the]] step S2917, the field information is reformed.

[0241] In the above (2), in the table, since all of the field structures of all of the records [[are]] coincide, the record can be picked up by matching the structure of the line with the field name with the field structure of the record. Further, the field is started from the field of the first record.

[0242] In [[a]] step S2929 shown in Fig. 29D, if the field exists, the program goes to [[a]] step S2920. If the field does not exist, the program goes to [[a]] step S2923. However, if no field exists at all, the processing of the multi-row/multi-column is ended.

[0243] In [[the]] step S2920, the structure of the field of one line is examined, and, in [[a]] step S2912, if the data of one line are all the same, since the table is a composite table, the processing of the multi-row/multi-column is ended.

[0244] Since it is necessary that the field structure of all of the records be coincided, in [[a]] step S2922, if the field structure of the field of one line examined up to now coincides with the structure of the field of one line examined in [[the]] step S2920, the program is returned to [[the]] step S2919. If it does not coincide, the program goes to [[a]] step ④.

[0245] In [[a]] step S2929, on the basis of the field structure of the record, the field information is reformed by matching the structure of the line with the field name with the field structure of the record.

[0246] In the above (3), since the table is a table in which the field portions of the field values are formed as the multi-row/multi-column, by re-reading the field portions having the multi-row/multi-column, the record can be picked up. Further, the field is started from the field of the first record.

[0247] In [[a]] step S2924 shown in Fig. 29E, if the field exists, the program goes to [[a]] step S2925. If the field does not exist, the processing of the multi-row/multi-column is ended.

[0248] In [[the]] step S2925, the structure of the field of one line is examined, and the program goes to [[a]] step S2926.

[0249] The fact that the field portion of the field value is more detailed means that this field includes the multi-row (or multi-column). Thus, in [[the]] step S2926, as a result that the structure of the field of one line is examined, if the structure is more detailed than the field name, the program goes to [[a]] step S2927. Otherwise, the processing of the multi-row/multi-column is ended.

[0250] In [[the]] step S2927, on the basis of the structure of the field of one line examined in [[the]] step S2925, the field information is reformed by matching the structure of the line with the field name with the field structure of the record.

[0251] As mentioned above, regarding the table to be processed, by analyzing the structure of the table and regularity of information description constituting the table and by reforming the table regularly in M columns and N rows, the table can be divided correctly.

[Thirteenth Embodiment]

[0252] In a thirteenth embodiment of the present invention, the HTML table reformation part 2501 effects treatment of the composite table. Here, on the basis of the table data stored in the HTML table storage part 101, by analyzing regularity of information description, the table is reformed regularly without omission of M columns and N rows.

[0253] The "composite table" is a table in which a plurality of tables are included in a single table and/or the record straddles between plural lines, so that the table analysis cannot be effected easily or simply.

[0254] The composite tables can be sorted into (1) a table in which the line with the field name is re-described in the table, (2) a table in which the same field names are included in series, (3) a table in which a field name (different from the common field name) and its field value are described on the way of the table, (4) a table in which a combination of adjacent tables is included in the table, and (5) others. Here, analyzing methods regarding the above (1) to (4) will be described.

[0255] Now, the details of the reformation of the HTML table in [[the]] step S2600 will be explained with reference to Figs. 31A to 31D.

[0256] Fig. 31A is a flow chart for processing the composite table in which the line with the field name is re-described in the table. Here, if the field name of each line with the field name is included in the record, such data is removed.

[0257] In [[a]] step S3101, a field name of one line is stored, and, in [[a]] step S3102, if the field exists, the program goes to [[a]] step S3103. If it does not exist, the program goes to ①.

[0258] In [[the]] step S3103, the field of one line is stored, and, in [[a]] step S3104, the fields of one line in [[the]] step S3101 is compared with that in [[the]] step S3103, and the program goes to [[a]] step S3105.

[0259] In [[the]] step S3105, as a result of comparison in [[the]] step S3104, if the fields are the same, the program goes to [[a]] step S3105, and, if not the same, in [[a]] step S3106, the field information is reformed.

[0260] Fig. 31B is a flow chart for processing the composite table in which the same field names are included in series. Here, when the field name of the line with the field name is described by plural times, arrangement of data is modified.

[0261] In [[a]] step S3107, if the field exists, the program goes to [[a]] step S3108. If the field does not exist, the program goes to [[a]] step S3112. However, if no field exists at all, the processing of the composite table is ended.

[0262] In [[the]] step S3108, one field name is stored, and the program goes to [[a]] step S3109. This field name is used for examining whether the same field name is described in the field name describing lines or not.

[0263] In [[the]] step S3109, all of the fields ~~all of the fields~~ of the lines with the field name are stored, and, in [[a]] step S3110, if the same field name exists in the lines with the field name, the program goes to [[a]] step S3111; whereas, if it does not exist, the program goes to ②.

[0264] In [[the]] step S3111, if the field names from lines regularly, the program is returned to [[the]] step S3107; whereas, if not, the program goes to ②.

[0265] In [[the]] step S3112, the reformation of the field information and reformation of positional relation graph are effected. For example, in Fig. 32A, the field names "000", "xxx" "△△△" form lines two times. Thus, by storing data of first series (portion shown by gray color) and then storing data of second series (portion shown by white color), the reformation is effected.

[0266] Fig. 31C shows a flow for processing a composite table in which field names different from the common field names and their field values exist on the way of the table. Here, when the field name describing lines in which the field means are changed partially are re-described and data for new lines with the field name are described in further fields, processing for correcting the order of data.

[0267] In [[a]] step S3113, a field name of a line is stored, and, in [[a]] step S3114, if the field exists, the program goes to [[a]] step S3115. If the field does not exist, the program goes to [[a]] step S3119. However, if no field exists at all, the processing of the composite table is ended.

[0268] In [[the]] step S3115, a field of a line is stored, and, in [[a]] step S3116, the fields of one line in [[the]] steps S3113 and S3115 are compared, and the program goes to [[a]] step S3117.

[0269] In [[the]] step S3117, as a result of comparison in [[the]] step S3116, if another field exists, the program goes to [[a]] step S3118; whereas, if another field does not exist, the program is returned to [[the]] step S3114.

[0270] In [[a]] step S3119, reformation of the field information and reformation of the positional relationship graph are effected.

[0271] For example, in Fig. 32B, there are field names "000", "xxx", "△△△" and "000", "□□□", "◎◎◎". Thus, the field names are made to "000", "xxx", "△△△", "□□□", ◎◎◎", and such data are stored and the reformation is performed.

[0272] Fig. 31D shows a flow of processing for a composite table in which there are a plurality of tables (lists) in the table. Here, when the field names are common and a plurality of tables are described in the single table, processing for dividing the tables or lists individually.

[0273] In [[a]] step S3120, a field name of a line is stored, and, in [[a]] step S3121, if the field exists, the program goes to [[a]] step S3122. If the field does not exist, the program goes to [[a]] step S3128. However, if no field exists at all, the processing of the composite table is ended.

[0274] In [[the]] step S3122, a field of a line is stored, and, in [[a]] step S3123, all of the fields stored in [[the]] step S3122 up to now are stored, and the program goes to [[a]] step S3124.

[0275] In [[the]] step S3124, if the same data exist in a line, since such data is a title, the program goes to [[a]] step S3125 to form a new table. If it does not exist, the program is returned to [[the]] step S3121. However, at a first time, the program does not go to [[the]] step S3125 but is returned to [[the]] step S3121.

[0276] In steps S3125 and S3126, objects of new field information object and new positional relationship are generated, and the program goes to [[a]] step S3127, where reformation of the field information is performed.

[0277] For example, in Fig. 32C, regarding the common field name, title 1 is described in a second line and title 2 is described in a fourth line. First of all, if there is the title 1 at the first time, since there is no data, a new table is not generated; if there is the title 2 at the second time, since the data regarding the title 1 has already been stored, a new table regarding the title 1 is generated. Lastly, if there is no field, since the data regarding the title 2 has already been stored, a new table regarding the title 2 is generated.

[0278] In [[a]] step S3128 and further steps, since the processing of the last title is not completed, post-treatment is performed.

[0279] First of all, in [[the]] step S3128, if the same data exist in a line, the program goes to [[a]] step S3129 to form a new table. If it does not exist, the processing of the composite table is ended.

[0280] In steps S3129 and S3130, objects of new field information object and new positional relationship are generated, and the program goes to [[a]] step S3131, where reformation of the field information is performed, and then the processing of the composite table is ended.

[0281] As mentioned above, regarding the table to be processed, by analyzing the structure of the table and regularity of information description constituting the table and by reforming the table regularly without omission of M columns and N rows, the table can be judged.

[Fourteenth Embodiment]

[0282] In a fourteenth embodiment of the present invention, the HTML table reformation part 2501 is constituted by a supplementary data removal part 3301 and a multi-column/multi-row processing [[art]] part 3302, as shown in Fig. 33.

[0283] Now, the details of the reformation of the HTML table in [[the]] step S2600 will be explained with reference to Fig. 34.

[0284] In [[a]] step S3401, supplementary data is removed from the HTML table, and, in [[a]] step S3402, by analyzing the structure of the table with reference to the table data from which the supplementary data is removed, the table is reformed regularly without omission of M columns and N rows, and the processing is ended.

[0285] As mentioned above, regarding the table to be processed, by analyzing the structure of the table and regularity of information description constituting the table and by reforming the table regularly without omission of M columns and N rows, the table can be judged.

[Fifteenth Embodiment]

[0286] In a fifteenth embodiment of the present invention, the HTML table reformation part 2501 is constituted by a supplementary data removal part 3501 and a composite table processing part 3502, as shown in Fig. 35.

[0287] Now, the details of the reformation of the HTML table in [[the]] step S2600 will be explained with reference to Fig. 36.

[0288] In [[a]] step S3601, supplementary data is removed from the HTML table, and, in [[a]] step S3602, by analyzing the regularity of the information description with reference to the table data from which the supplementary data is removed, the table is reformed regularly without omission of M columns and N rows, and the processing is ended.

[0289] As mentioned above, regarding the table to be processed, by analyzing the structure of the table and regularity of information description constituting the table and by reforming the table regularly without omission of M columns and N rows, the table can be judged.

[Sixteenth Embodiment]

[0290] In a sixteenth embodiment of the present invention, the HTML table reformation part 2501 is constituted by a supplementary data removal part 3701, a multi-column/multi-row processing part 3702 and a composite table processing part 3703, as shown in Fig. 37.

[0291] Now, the details of the reformation of the HTML table in [[the]] step S2600 will be explained with reference to Fig. 38. In [[a]] step S3801, supplementary data is removed from the HTML table, and, in [[a]] step S3802, by analyzing the structure of the table with reference to the table data from which the supplementary data is removed, the table is reformed regularly without omission of M columns and N rows, and the program goes to [[a]] step S3803.

[0292] In [[the]] step S3803, by analyzing the regularity of information description with reference to the reformation data of [[the]] step S3802, the table is reformed regularly without omission of M columns and N rows, and the processing is ended.

[0293] As mentioned above, regarding the table to be processed, by analyzing the structure of the table and regularity of information description constituting the table and by reforming the table regularly without omission of M columns and N rows, the table can be judged.

[Seventeenth Embodiment]

[0294] In a seventeenth embodiment of the present invention, the HTML table reformation part 2501 is constituted by a multi-column/multi-row processing part 3901 and a composite table processing [[art]] part 3902, as shown in Fig. 39.

[0295] Now, the details of the reformation of the HTML table in [[the]] step S2600 will be explained with reference to Fig. 40.

[0296] In [[a]] step S4001, by analyzing the structure of the table with reference to the table data from which the supplementary data is removed, the table is reformed regularly without omission of M columns and N rows, and the program goes to [[a]] step S4002.

[0297] In [[the]] step S4002, by analyzing the regularity of information description with reference to the reformation data of [[the]] step S4001, the table is reformed regularly without omission of M columns and N rows, and the processing is ended.

[0298] As mentioned above, regarding the table to be processed, by analyzing the structure of the table and regularity of information description constituting the table and by reforming the table regularly without omission of M columns and N rows, the table type can be judged.

[0299] Incidentally, the present invention may be applied to a system including a plurality of units of equipment[[s]] (for example, a computer body, an interface equipment, a display and the like) or a system including a single piece of equipment, so long as the functions of the above-mentioned embodiments can be realized.

[0300] Further, a technique in which, for the purpose for operating various devices to realize functions of the above-mentioned embodiments, software program code for realizing functions of the above-mentioned embodiments is supplied to a computer (or CPU or MPU) in an apparatus or a system connected to various devices so that the various devices are operated by the computer in the apparatus or the system in accordance with the program code, is also included within the scope of the invention. Further, in this case, the program code itself read out from a recording medium realizes the functions of the above-mentioned embodiments, and, thus, the program code itself and means for supplying the program code to the computer (for example, recording medium storing the program code) constitute the present invention.

[0301] The recording medium for supplying the program code may be, for example, a floppy disk, a hard disk, an optical disk, a photo-magnetic disk, CD-ROM, CD-R, a magnetic tape, a non-volatile memory card or ROM.

[0302] Further, when not only the functions of the above-mentioned embodiments are realized by carrying out the program code read out from the computer but also the functions of the above-mentioned embodiments are realized by cooperation with OS (operating system) operating on the computer or other application software on the basis of instruction of the program code, such program code is included within the scope of the invention.

[0303] Further, of course, the present invention includes a technique in which, after the program code read out from the recording medium is written in a memory of a function expansion board inserted into the computer or a function expansion unit connected to the computer, a CPU of the function expansion board or the function expansion unit carries out the actual processing partially or totally on the basis of instruction of the program code, thereby realizing the functions of the above-mentioned embodiments.

[0304] When the present invention is applied to the above-mentioned recording medium, program codes corresponding the above-mentioned flow charts may be stored in the recording medium.

[0305] Although the present invention has been described in its preferred forms with a certain degree of particularity, many apparently widely different embodiments of the invention can be made without departing from the spirit and the scope thereof. It is to be understood that the invention is not limited to the specific embodiments thereof except as defined in the appended claims.